# Discovering cognitive models in a competitive mixed-strategy game

**Peiyu Liu (peiyu.liu@yale.edu)**
Department of Psychiatry, Yale University School of Medicine
New Haven, CT 06510, US

**Kevin J. Miller (kevinjmiller@deepmind.com)***
Google DeepMind and University College London
London, United Kingdom

**Hyojung Seo (hyojung.seo@yale.edu)***
Department of Psychiatry, Yale University School of Medicine
New Haven, CT 06510, US

* co-corresponding author

## Abstract:

**Sophisticated behavioral tasks are key tools in cognitive neuroscience, but pose challenges because the cognitive processes that give rise to behavior are often incompletely understood. Matching pennies (MP) is one such task: a strategic zero-sum game that has been broadly used for theoretical and empirical analysis of dynamic social interactions across species. Disentangled recurrent neural networks (disRNN) are a recently introduced deep learning method which allows discovering cognitive hypotheses directly from behavioral datasets. Here, we apply disRNN to a widely studied dataset of non-human primates playing an iterative MP game. We find that the discovered models provide a better qualitative and quantitative match to behavior than classic behavioral models, and reveal readily-interpretable cognitive hypotheses. Specifically, they show that animal's behavior can be described as a mixture of long-term heuristics such as choice perseveration and reward-following, as well as short-term strategies that contribute to countering the opponent's strategy.**



**Keywords: matching pennies (MP); reinforcement learning; behavioral models; disentangled RNN (disRNN)**


## Introduction

Matching pennies (MP) is a zero-sum game, for which an equilibrium or optimal strategy from the game-theoretic perspective is for each player to randomize his/her choice, and make it unpredictable to the opponent (von Neumann and Morgenstern, 1944). When such a mixed strategy game is iteratively played, humans dynamically change their choice as they build their beliefs about the opponent's strategy and also infer the opponent's beliefs on their own strategy over the course of the game (Camerer, 2003; Hampton et al., 2008). Iterative MP has been adopted to investigate neural mechanisms of dynamic and strategic decision-making in animal models. When pitted against a computerized opponent that exploited and punished any serial dependencies in choice and reward within relatively short sequences spanning a few trials, animals were able to randomize their choice or even counter the opponent's exploitative strategy deviating from the equilibrium strategy (Lee et al., 2004; Seo et al., 2014; Tervo et al., 2014).

Choice behavior of rhesus monkeys typically showed a mixture of a serial correlation with self- and opponent's past choice that slowly decayed over several trials, and an abrupt reduction/removal of the dependency of a choice on the choice and reward from the immediately preceding trial (Fig. 1, a, c). Therefore, a best human-derived cognitive model, namely a *forgetting Q-learning model (fQ)* only partially explained animal's behavior, failing to capture the full complexity of the dynamic behavior during the game (Barraclough et al., 2004; Ito et al., 2009).

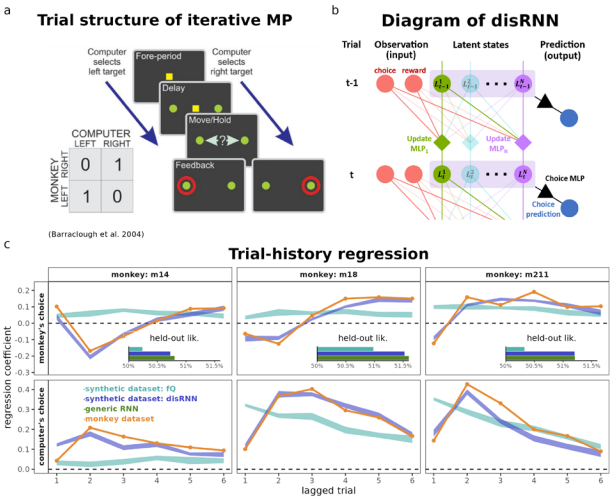


**Figure 1**. (a) Trial structure of an iterative MP. (b) Schematic diagram of disRNN architecture. (c) Serial correlation of a choice with the choices of self (top) and the opponent (bottom) over past trials compared with synthetic data from models. Inset: quality-of-fit of best fQ, disRNN and generic RNN model.

To discover internal/cognitive strategies that can give rise to the observed behavioral phenomena, here we adopted a recently developed disentangled recurrent neural network (disRNN; Miller et al., 2023). disRNN encourages the network to learn sparse representations in which each dimension (latent state) corresponds to a single factor of variation in the data ("disentangling") by implementing the update rule of each latent state in a separate module of sub-networks (multilayer Perceptrons, MLPs), and by using information bottlenecks imposing a penalty on maintaining information within the network and sub-networks (Fig. 1, b).

For each animal, disRNN identified multiple separable strategies - long-term heuristics that track/repeat self- and the opponent's choice, mixed with short-term strategies that reduce serial correlation produced by these heuristics to evade potential exploitation by the opponent. Our results demonstrate that disRNN can successfully discover readily interpretable cognitive strategies intertwined together to generate complex behavioral phenomena as in a mixed strategy game.

## Methods

Behavioral task. Iterative MP was implemented as a binary oculomotor choice task, in which animals chose between two targets presented at the left and right side of a central fixation target by shifting their gaze (Barraclough et al., 2004; Fig. 1, a). The opponent's choice was indicated by a red ring around the chosen target. Animals received juice reward only if their choice matched the opponent's.

Modeling. We used a behavioral dataset collected from three monkeys (Kim et al., 2007). In forgetting Q-

learning model, choice-specific reward expectation or action value was learned through an iterative update as follows: $Q_{t+1}(A) = \alpha_F \cdot Q_t(A) + \Delta_i$, where $\alpha_F$ is a forgetting/decay rate for an old estimate, $\Delta_i$ is an incremental change separately after reward ($i$ =1) and no-reward ($i$ =0), and $A$ is action. A softmax function was used for action selection with the value difference between two actions as its input. Parameters were estimated to maximize the likelihood of held-out data.

disRNN models were implemented as described in (Miller et al. 2023). Best model for each monkey was selected to maximize the likelihood of held-out data across varying hyperparameters such as the size of sub-networks, recurrent network, and penalty scale. To examine the qualitative fit of the best disRNN model, we simulated choice by pitting each model against the computerized opponent using exploitative algorithms identical to those used against monkeys during MP.

## Results

A best-fitting disRNN model typically comprised a small number of recurrent (latent) states updated at each time step by simple latent-specific rules that integrate observations and latent states from the previous time step through open information bottlenecks (Fig. 1, b). These models showed a good quality of fit and recapitulated individual patterns of serial correlation in choice/reward better than the forgetting Q-learning model (Fig. 1, c).

Information processing of each latent state ($L_t^i$) was interpretable by the following update rule:

$$L_{t+1}^i = L_t^i + \alpha_t^i \cdot \left(R_t^i - L_t^i\right) = \left(1 - \alpha_t^i\right) \cdot L_t^i + \alpha_t^i \cdot R_t^i$$

where $R_t^i$ is an incremental change (i.e. intercepts in Fig. 2, c-f). $\alpha_t^i$ is a learning rate – conversely, $1 - \alpha_t^i$ is a decay rate (i.e. slopes in Fig. 2, c-f) that determines how fast/slow the latent state is updated with a new increment at a given time step. By inspecting how these two parameters were determined by observations and other latent states, we could analyze what information is integrated (processed) through each latent state. We will use a specific model discovered for one animal (m14) to illustrate this point (Fig. 2).

disRNN identified 4 latent states, each of which potentially corresponds to a single strategic component. $L^1$ changed by positive (negative) $R_t^1$ after right (left) choice with a small learning rate $\alpha_t^i$, thereby reflecting self-choice accumulated over many trials (Fig. 2, c). $L^2$ changed by positive (negative) increments $R_t^2$, after left (right) choice was rewarded or right (left) choice was not rewarded, thereby accumulating *choice-specific reward* (equivalently, *the opponent's choice*) with a small $\alpha^2$ (Fig. 2, d). $L^1$and $L^2$ appeared to capture the animal's tendency of tracking the side frequently chosen by self and the opponent over many trials in the past. By contrast, $L^3$ accumulated self-choice with a relatively

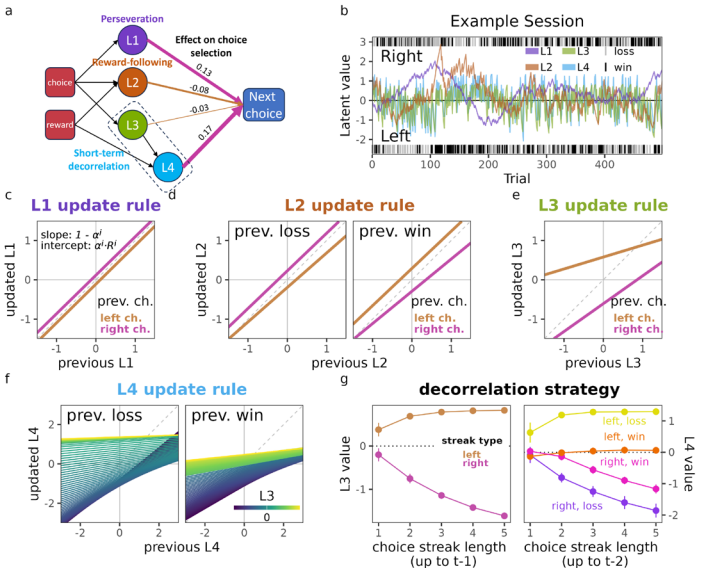


**Figure 2**: Cognitive strategy discovered by DisRNN. (a) Dependency graph of the discovered model. (b) DisRNN run with the choices and rewards from an example behavioral session. (c-f) Visualization of learned update rules. (g) choice decorrelation through $L^3, L^4$ modulation dependent on a recent streak of repetitive choice.

large learning rate (Fig. 2, e). It reset toward 0 after the animal switched choice, and then rapidly increased (decreased) with repetitive left (right) choice (Fig. 2, g).

Interestingly, $L^4$ was conjointly modulated by $L^3$ and previous reward, likely to reflect a higher-order strategy (Fig. 2, f). When $R_t^4$ inversely correlated with $L^3$, overall large $\alpha^4$ and its effect on the subsequent choice are taken all together, $L^4$ was apparently monitoring repetitive choice in recent trials and contributing to rapidly switching side, with such a tendency particularly stronger when the previous choice failed to obtain a reward than it was when rewarded (Fig. 2, g).

Could the cognitive strategies identified by disRNN give the animal a leverage to win the game?
We find that the animal's tendency to temporally intermix mutually antagonistic strategies may provide an adaptive solution for mixed needs – a need for a effort-efficient long-term strategy on the one hand to generate a long series of choice to play an iterative game, and a need for short-term strategies on the other hand to add variability and decorrelate adjoining choices, countering the opponent's exploitation of serial correlations produced by the long-term strategies.

## Future Directions & Acknowledgements

We plan to design quantitative behavioral analyses, and systematic perturbations of latent states to further test and validate the cognitive models and strategic components discovered by disRNN.

This work was supported by NIH R01NS118463.

## References


Barraclough, D. J., Conroy, M. L., & Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. Nature Neuroscience, 7(4), 404–410. https://doi.org/10.1038/nn1209

Lee, D., Conroy, M. L., McGreevy, B. P., & Barraclough, D. J. (2004). Reinforcement learning and decision making in monkeys during a competitive game. Cognitive Brain Research, 22(1), 45–58. https://doi.org/10.1016/j.cogbrainres.2004.07.007

Miller, K. J., Eckstein, M., Botvinick, M. M., & Kurth-Nelson, Z. (2023). Cognitive Model Discovery via Disentangled RNNs (p. 2023.06.23.546250). bioRxiv. https://doi.org/10.1101/2023.06.23.546250

Tervo, D. G. R., Proskurin, M., Manakov, M., Kabra, M., Vollmer, A., Branson, K., & Karpova, A. Y. (2014). Behavioral Variability through Stochastic Choice and Its Gating by Anterior Cingulate Cortex. Cell, 159(1), 21–32. https://doi.org/10.1016/j.cell.2014.08.037

Von Neumann, J. and Morgenstern, O. (1944). Theory of games and economic behavior. Princeton University Press. Princeton, NJ.

Camerer F.C. (2003). Behavioral game theory: experiments in strategic interaction. Princeton University Press. Princeton, NJ.

Hampton, A.N., Bossaerts, P., O'doherty, J.P. (2008). Neural correlates of metalizing-related computations during strategic interactions in humans. PNAS, 105: 6741-46.

Seo, H., Cai, X., Donahue, C.H., Lee, D. (2014). Neural correlates of strategic reasoning during competitive games. Science, 346: 340-3.

Kim, S., Hwang, J., Seo, H., Lee, D. (2009). Valuation of uncertain and delayed rewards in primate prefrontal cortex. Neural networks, 22:294-304.